 ISBN 978-80-7378-224-5

Image Quality Assessment

M. Kudělka Jr.
Institute of Information Theory and Automation,
Academy of Sciences of the Czech Republic, Prague, Czech Republic.

Abstract. Image quality assessment is a difficult problem in the field of image processing without any acceptable solution yet. Quality assessment of textures, which is the topic of this work, is even harder problem. This work presents an overview and a description of current state of the art methods as well as brief experiments on textures and a discussion about the usability of these methods on textures.

Introduction

Image quality assessment tries to quantify a visual quality or, analogically, an amount of distortion in a given picture. These distortions are inevitable part of any digital image processing pipeline (acquisition, compression, transmission, etc. of images). The only "correct" method of evaluating the human-perceived visual quality of the pictures is the evaluation by the human beings. Unfortunately, such a procedure is expensive, very time consuming and not usable in real-time applications (adjustment of the rate of transfer etc.). Therefore, there is a need for an automated method that would predict the human-perceived visual quality as close as possible.

The pioneering work in this area was done by psychologist and neuroscientist Béla Julesz. In the article Visual Pattern Discrimination from 1962 [Julesz, 1962] he experimented with textures with controlled properties to find out what is important for a human to discriminate two textures. The age of this article shows how old the problem is and also how difficult it is, because it is still not satisfyingly solved.

This work is oriented on an evaluation of texture quality, which has not been very well documented yet. Its task differs from regular quality assessment. The ideal measure would predict and quantify how much the tested (e.g. synthesized) texture can or cannot be distinguished from the original texture by a human. This cannot be achieved by any kind of a pixel-wise comparison, because, usually, the exact pixel-to-pixel correspondence is not necessary or is even undesired. Analysis of its structure or its statistical properties could be the right approach.

Most of today's methods are designed to evaluate the visual quality of real-world images like photos. Within this work the survey of most used methods in this area was made to see if they or at least some ideas from them could be refined and used. All of studied methods work with monospectral (greyscale) images. They could also be used with multispectral images after the preprocessing, but this leads to information loss. In the next part there are presented possible classifications of image quality measures and then the measures themselves are described.

Classifications

Most basically, image quality measures can be divided to *subjective* and *objective* ones. The former are performed by humans, i.e. the image quality is evaluated by humans; the latter are acquired by computer algorithms.

Another possible classification of image quality measures can be made according to the availability of a reference image. Most of existing approaches are *full-reference* (FR), which means that complete reference image is available during an evaluation. Reference image is often not available at all and *no-reference* (NR) or "blind" methods are needed. Third group is called *reduced-reference* (RR) and such methods measure image quality with help of features extracted from a reference image. The measures considered in this work are objective and belong to the FR class.

MSE and PSNR

Mean-squared error (MSE) and peak signal-to-noise ratio (PSNR) are two basic measures. They measure a difference between two signals and the result can be understood as a degree of similarity or a strength of error signal between signals.

Let $\mathbf{x}$ and $\mathbf{y}$ be two discrete image signals (generally any signals) of finite length N, where N is the number of pixels (samples) in the images and x_i and y_i are intensities of i-th pixel in $\mathbf{x}$ and $\mathbf{y}$, respectively. Then MSE between these two signals is

$$\mathrm{MSE}(\mathbf{x}, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} (x_i - y_i)^2. \tag{1}$$

In the field of image processing the MSE is often converted into PSNR measure

$$\mathrm{PSNR}(\mathbf{x}, \mathbf{y}) = 10 \log_{10} \frac{L^2}{MSE(\mathbf{x}, \mathbf{y})}, \tag{2}$$

where L is a dynamic range of allowable pixel intensities. For example if the image has 8 bits for a pixel, $L = 2^8 - 1 = 255$. PSNR is useful when comparing images with different dynamic ranges [Wang and Bovik, 2009].

This measure is universal, easily computable and as a valid distance metric in $\mathbb{R}^N$ has a few nice conditions such as symmetry, triangular equality etc. and therefore provides consistent interpretation of image similarity. Because of all of this MSE became a convention in image quality evaluation and has been compared to new methods in this field.

So what is the problem with MSE? It measures just the pixel-wise correspondence and the main issue, as many experiments and tests of MSE have shown, is that despite of all the good features, MSE does not represent human-perceived image quality very well. This led to attempts to create measures, whose performance would be more closely related to the human perception of visual quality.

Modeling of human visual system

Because the difference between signals does not measure the distortion well, it is natural to try and model human visual system (HVS) itself. If all parts of HVS were precisely modeled, accurate prediction of subjective image visual quality would probably be achieved. Precise modeling of HVS is, however, very hard, if not impossible, because HVS is very complicated system with a lot of nonlinearities and, moreover, we still do not know every detail of how it processes information.

In the pioneer method of this approach [Mannos and Sakrison, 1974] the HVS is modeled by monotonic, increasing, concave function $f(i)$ and a linear filter $A(f_x, f_y)$. Let $u(x, y)$ be an intensity of the pixel $[x, y]$ of an image. First, on both reference and distorted images the function $f(i)$ is applied and a new image $w(x, y) = f(u(x, y))$ is created. Then, the filter $A(f_x, f_y)$ is applied to obtain the final transformed image $v(x, y) = FT^{-1}\left(A(f_x, f_y) FT\left(w(f_x, f_y)\right)\right)$. Finally, the distortion d is measured by the integral squared error $d(v, v') = \int_x \int_y \left(v(x, y) - v'(x, y)\right)^2 dxdy$.

This model should at least partially correspond to image processing of human visual system. Function f represents the sensitivity of an eye to the light and the filter A has some correspondence to a lateral inhibition and optical limitations of human visual system. However, this method and others similar to it are essentially just variants of MSE, only different parts of the signals are weighted differently according to their presumed visibility for the HVS.

SSIM

Structural similarity index is a measure based on the assumption that human visual system is adapted to extract structural information from the field of view. Therefore, the change of

structural information between distorted and original image could be a good approximation of perceived image distortion.

In [Wang et al., 2004] basic version of SSIM is described, where structural information is gathered by a comparison of luminance, contrast and structure. Let $\mathbf{x}$ and $\mathbf{y}$ are two nonnegative image signals of length N. First, the luminance is compared by a function of mean intensities μ_x and μ_y in the form

$$l(\mathbf{x}, \mathbf{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, \tag{3}$$

where C_1 is the constant included to avoid an instability when $\mu_x^2 + \mu_y^2$ is almost zero. $C_1 = (K_1 L)^2$, where L is a dynamic range of the picture and $K_1 << 1$ is a small number. The same stands for the constants in contrast and structure functions, whose description follows.

Contrast comparison is a function of standard deviations σ_x and σ_y that looks like

$$c(\mathbf{x}, \mathbf{y}) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \tag{4}$$

where $C_2 = (K_2 L)^2$ and $K_2 << 1$. Also note, that with the same amount of change of contrast $\Delta\sigma = \sigma_y - \sigma_y$, this function is less sensitive in the case of high base contrast than in the case of low base contrast, which corresponds to the contrast-masking behaviour of human visual system. Finally, structure comparison is a function of correlation between the two signals in the form

$$s(\mathbf{x}, \mathbf{y}) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}, \tag{5}$$

where C_3 is again a small constant. Note, that correlation coefficient between orignal signals is the same as between normalized signals (i.e. $(\mathbf{x} - \mu_x)/\sigma_x$) and therefore could represent the structure well.

These functions are combined into the resulting measure as

$$\mathrm{SSIM}(\mathbf{x}, \mathbf{y}) = [l(\mathbf{x}, \mathbf{y})]^\alpha \, [c(\mathbf{x}, \mathbf{y})]^\beta \, [s(\mathbf{x}, \mathbf{y})]^\gamma, \tag{6}$$

where $\alpha > 0$, $\beta > 0$, $\gamma > 0$ are parameters used to change the relative importance of individual components. For simplification in [Wang et al., 2004] they set $\alpha = \beta = \gamma = 1$ and $C_3 = C_2/2$, which simplified the measure to

$$\mathrm{SSIM}(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \tag{7}$$

SSIM works best if used locally. It means to compute local statistics μ_x, σ_x and σ_{xy} in a small window that is pixel-by-pixel moved over the entire image and the results are then averaged. The reasons for such an approach are that different parts of the image can differ a lot and also human can concentrate on just one limited area at the time. Such approach can also be used to create spatially varying quality map of the picture to obtain more information about the distortion of the image.

SSIM is one of the most used measures not only in the field of image processing. For example, SSIM is used in award-winning freeware H.264 codec x.264 and it is also used in speech recognition, in compressing algorithms etc.

SSIM, even though it performs much better than MSE, has limits. For example, the basic variant does not perform well in cases of translated, scaled or rotated images, even if the quality of these images is the same as of their reference images. This is partially solved by Complex Wavelet SSIM (CW-SSIM) [Wang and Simoncelli, 2005]. SSIM, in essence, also compares the signals with pixel-to-pixel approach so it is still quite similar to MSE.

VIF criterion

Main theoretical problem with SSIM is the assumption about the structural information. There is no widely accepted definition of this term and, therefore, such an assumption does not have to be correct. Another attempt of image quality assessment is *visual information fidelity* (VIF) criterion presented in [Sheikh and Bovik, 2006].

Authors of VIF also assumes that HVS has evolved to best perceive so called natural scenes. However, instead of HVS, they have chosen to model these natural scenes, which are a class of images or videos of common three-dimensional visual environment. Even though this class forms just a small subspace in the space of all possible signals, its relevance for a human caused an increased interest in a statistics of their structure [Srivastava et al., 2003], which are called natural scenes statistics (NSS).

They model the reference image as an output of a natural source that passes through the HVS channel (signal for a test image passes also through the distortion channel). Then, the amount of information brain could extract is said to be the mutual information between the output of the source and the HVS channel. Everything is modeled for a single subband in the wavelet domain, which was shown to well represent the natural sources. In the final VIF measure these values for both reference and test image are compared.

Natural source is modeled as Gaussian scale mixtures (GSM), which is a random field (RF) that can be expressed as a product of two independent RFs. That is, $\mathcal{C} = \mathcal{S} \cdot \mathcal{U} = \{S_i \cdot \overrightarrow{U_i} : i \in I\}$, where I is a set of spatial indices, $\mathcal{S} = \{S_i : i \in I\}$ is an RF of positive scalars and $\mathcal{U} = \{\overrightarrow{U_i} : i \in I\}$ is a Gaussian vector RF with zero mean and covariance $\mathbf{C}_U$. Subbands in the wavelet domain are split into nonoverlapping blocks of M coefficients and the block i is modeled as the M-dimensional vector $\overrightarrow{C_i}$.

Distortion channel is modeled as a signal attenuation plus additive noise. This model is supposed to well approximate real-world distortions locally and is not specialized to particular artifacts (e.g. blocking of JPEG compression). The model of the output of the distortion channel looks like $\mathcal{D} = \mathcal{G} \cdot \mathcal{C} + \mathcal{V} = \{g_i \overrightarrow{C_i} + \overrightarrow{V_i} : i \in I\}$, where RF $\mathcal{G} = \{g_i : i \in I\}$ is a deterministic scalar gain field and $\mathcal{V} = \{\overrightarrow{V_i} : i \in I\}$ is a stationary, additive, zero-mean, white Gaussian noise RF with variance $\mathbf{C}_V = \sigma_v^2 \mathbf{I}$.

HVS channel is modeled simply as a single additive noise component that adds uncertainty to the signal that flows through the HVS. It is again a stationary, zero mean, additive, white Gaussian noise RF $\mathcal{N} = \{\overrightarrow{N_i} : i \in I\}$ for a reference and $\mathcal{N}'$ for a test image. Then $\mathcal{E} = \mathcal{C} + \mathcal{N}$ and $\mathcal{F} = \mathcal{D} + \mathcal{N}'$ are the signals processed by the brain for a reference and a test image, respectively.

From this model, the mutual information for both images $I(\overrightarrow{C}, \overrightarrow{F}|s)$ and $I(\overrightarrow{C}, \overrightarrow{E}|s)$ can be computed, where s stands for the realization of S for particular reference image (detailed description of the calculation of mutual information can be found in [Sheikh and Bovik, 2006]).

Final measure then combines these results from each subband into the final formula

$$\mathrm{VIF} = \frac{\sum_{j \in \mathrm{subbands}} I(\overrightarrow{C}^{N,j}, \overrightarrow{F}^{N,j}|s^{N,j})}{\sum_{j \in \mathrm{subbands}} I(\overrightarrow{C}^{N,j}, \overrightarrow{E}^{N,j}|s^{N,j})}, \tag{8}$$

where $\overrightarrow{C}^{N,j}$ (and others analogically) represents the N elements of RF $\mathcal{C}_j$ for j-th subband. This N-notation is useful for a local application of the measure, because as well as SSIM, VIF works better if computed locally by a moving window.

For common distortions values of the VIF belong to the interval $[0, 1]$, where VIF $= 1$ if and only if the test image is the copy of the reference image and VIF $= 0$ if all information was lost because of the distortion. However, for the special case of slight linear contrast enhancement of the test image the value of VIF will be greater than 1. This is quite useful property, because such images are visually perceived to be better than originals. According to [Sheikh et al., 2006], VIF criterion performs better than all other state-of-the-art image quality assessment methods.

Experiments with textures

Tested measures were MSE, VSNR (Visual Signal-to-Noise-Ratio [Chandler and Hemami, 2007], which is based on HVS modeling), SSIM and VIF. The implementations of the aforementioned methods from the Matlab package MeTriX_MuX [Gaubatz, 2012] were used.

Figure 1. Textures of wood. From the left: original, 2D GMRF, 3D GMRF, CAR, SAR.

All of the measures were used on two sets of textures: wood and straws. Textures used were modeled with 2D Gaussian Mixtures Random Field (GMRF), 3D GMRF, Causal Auto-Regressive model (CAR) and Spatial Auto-Regressive model (SAR) [Haindl, 1991]. Lets label them with numbers from 1 to 4 in this order. An example of a texture of wood can be seen on Figure 1.

Table 1. Test results on generated textures of wood.

	2D	3D	CAR	SAR
mse	383.943825	683.548720	605.675215	472.032387
vsnr	4.154531	0.685098	1.493072	2.847859
ssim	0.423054	0.258032	0.292863	0.293117
vif	0.019130	0.020040	0.026589	0.015571

The results on a wood textures can be seen in the Table 1. Lets look at the textures themselves. Their visual quality is a matter of opinion, but the discussion will be as objective as possible. The SAR generated texture (the last one on the Figure 1) has a bad structure and resembles the original wood texture just by its color. Therefore, it should be evaluated as the worst one. The 3D GMRF generated texture is quite similar to the original. Even though its structure is not exactly the same, it looks a lot like the original wood and the distribution of colors in the image is also fairly similar to the original. Therefore, it should be evaluated as the best one. The 2D GMRF and CAR generated textures are both completely different and it is hard to compare them. The 2D GMRF has very soft structure that may look like a wood, but not as a wood in the original image. The CAR, on the other hand, has very aggressive texture but is slightly similar to the original. However, both of them, for sure, are better than SAR and worse than 3D GMRF. So, the textures should be evaluated from the best to the worst in the order 2, 1, 3, 4 or 2, 3, 1, 4.

According to the results, MSE evaluates the textures as 1, 4, 3, 2, VSNR evaluates them as 2, 3, 4, 1, SSIM also as 2, 3, 4, 1 and VIF as 4, 1, 2, 3. MSE is completely wrong, which was

expected due to its poor perfomance on normal images. VSNR and SSIM evaluated textures the same, which could also be expected because of their common pixel-to-pixel approach. They both valued the second texture correctly as the best one, but SSIM just by a small margin above others and both of them did not evaluated correctly the rest. VIF performed surprisingly bad, even though the final ordering is slightly better than the ordering of MSE. Textures from the second set has poor visual quality and the tests on them turned out even worse and therefore are not mentioned here.

Conclusion

Image quality assessment is an important problem in the field of image processing. It is still not satisfyingly solved and new approaches are still appearing. This work concentrates on a texture quality assessment and tries to find out whether the current state of the art methods for images can be used on textures. The most commonly used measures for image quality assessment were described and briefly tested on textures. The tests showed that none of the tested methods works well on used textures (not even SSIM or VIF) and, therefore, there is a need for a measure designed particularly for textures, which is the subject of further effort. Important disadvantage of the most of the current methods is also that they are designed to work only with grey-scale images. The use of information from all scales could help with the quality assessment. There should be also more thorough testing on the variety of textures to show with greater certainity that the methods for images do not work on textures or find the cases where they do work and determine why.

References

Chandler, D. and Hemami, S., Vsnr: A wavelet-based visual signal-to-noise ratio for natural images, *IEEE Trans. on Image Processing*, *16*, 2284–2298, 2007.

Gaubatz, M., Metrix_mux, URL `http://foulard.ece.cornell.edu/gaubatz/metrix_mux/`, 2012.

Haindl, M., Texture synthesis, *CWI Quarterly*, *4*, 305–331, 1991.

Julesz, B., Visual pattern discrimination, *IRE Trans. on Information Theory*, *8*, 84–92, 1962.

Mannos, J. and Sakrison, D., The effects of a visual fidelity criterion of the encoding of images, *IEEE Transactions on Information Theory*, *20*, 525–536, 1974.

Sheikh, H. and Bovik, A., Image information and visual quality, *Image Processing, IEEE Transactions on*, *15*, 430–444, 2006.

Sheikh, H., Sabir, M., and Bovik, A., A statistical evaluation of recent full reference image quality assessment algorithms, *IEEE Trans. on Image Processing*, *15*, 3440–3451, 2006.

Srivastava, A., Lee, A., Simoncelli, E., and Zhu, S., On advances in statistical modeling of natural images, *Journal of mathematical imaging and vision*, *18*, 17–33, 2003.

Wang, Z. and Bovik, A., Mean squared error: Love it or leave it? a new look at signal fidelity measures, *Signal Processing Magazine, IEEE*, *26*, 98–117, 2009.

Wang, Z. and Simoncelli, E., Translation insensitive image similarity in complex wavelet domain, in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 2, pp. 573–576, IEEE, 2005.

Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E., Image quality assessment: From error visibility to structural similarity, *Image Processing, IEEE Transactions on*, *13*, 600–612, 2004.